ACPVI multibloc qualitative. Application en épidémiologie vétérinaire.

Stéphanie BOUGEARD [1], El Mostafa QANNARI [2] & Christelle FABLET [1]

(1) *AFSSA, Département d'épidémiologie animale - Zoopôle, BP53, 22 440 Ploufragan*
(2) *ENITIAA-INRA, Unité de Sensométrie et Chimiométrie - Rue de la Géraudière BP 82225, 44322 Nantes Cedex*

Résumé

Une nouvelle présentation de l'Analyse en Composantes Principales sur Variables Instrumentales Multibloc, dont l'objectif est de prédire un tableau Y à partir de plusieurs tableaux $(X_1, \ldots, X_K)$, est proposée. Elle est basée sur la détermination, pas à pas, de composantes dans l'espace des variables Y. Chaque composante est projetée sur les espaces engendrés respectivement par les variables des tableaux X_k $(k = 1, \ldots, K)$. L'$ACPVI$ multibloc consiste à maximiser, en moyenne, la variance restituée par ces projections. Cette méthode multibloc est ensuite appliquée au cadre de la description et la prédiction d'une variable qualitative y par un ensemble de variables qualitatives $(x_1, \ldots, x_K)$, chaque variable étant codée en un tableau contenant les indicatrices de ses modalités. La discrimination est opérée sur la base de composantes globales mutuellement orthogonales résumant l'ensemble des variables explicatives. La démarche d'analyse est comparée à d'autres méthodes de discrimination qualitative et illustrée sur la base d'une étude de cas en épidémiologie vétérinaire.

Abstract

A new presentation of Multibloc Redundancy Analysis is discussed. This method of analysis aims at predicting a set of variables Y from several blocks of variables $(X_1, \ldots, X_K)$. It consists in determining, step by step, components in the Y space which are projected upon the spaces spanned by the variables in X_k $(k = 1, \ldots, K)$. At each step, the component is sought in such a way so as to maximize the averaged variance explained by the projections. Thereafter, the method of analysis is applied to the case of categorical variables, each variable being coded by the indicators of its categories. The discrimination and classification is achieved using orthogonal components from the predictive variables. We also outline how the method is related to other qualitative discriminant techniques. The interest of the method is illustrated on the basis of a case study in the field of veterinary epidemiology.

Mots clés

Analyse canonique généralisée, analyse en composantes principales sur variables instrumentales multibloc, analyse des correspondances multiples avec un tableau de référence, discrimination qualitative.

1 Introduction

L'explication d'une variable y par un ensemble de variables $(x_1, \ldots, x_K)$, toutes qualitatives, est une problématique statistique classique. Du fait du grand nombre de variables explicatives et des liaisons structurelles des données d'épidémiologie vétérinaire, nous choisissons de nous situer dans le cadre des méthodes factorielles. L'$ACPVI$ multibloc (Bougeard *et al.*, 2007) est initialement développée pour le traitement de données quantitatives organisées en $(K+1)$ tableaux. Elle peut être adaptée au cadre qualitatif en considérant que chaque bloc est constitué par l'ensemble des modalités d'une variable qualitative codée de façon disjonctive. La détermination de composantes globales mutuellement orthogonales permet d'optimiser le modèle de discrimination qualitative issu de cette méthode.

2 ACPVI multibloc

Dans une publication précédente, nous avons présenté l'$ACPVI$ multibloc en nous basant sur plusieurs critères qui reflètent différentes facettes de cette méthode (Bougeard *et al.*, 2007). Nous présentons ici une nouvelle façon d'introduire cette méthode qui souligne notamment ses liens avec l'analyse canonique généralisée, ACG (Carroll, 1968).

Nous disposons d'un tableau de variables quantitatives X partitionné en K blocs $X = [X_1|\ldots|X_K]$. Chaque tableau X_k est un tableau $(N \times P_k)$ dont les lignes correspondent aux mêmes N individus. Toutes ces variables sont supposées centrées. On désigne par la suite par $P_{X_k} = X_k(X_k'X_k)^{-1}X_k'$ le projecteur associé au sous-espace engendré par les variables de X_k. L'ACG peut être définie comme une démarche pas à pas, qui consiste à déterminer à chaque étape une variable canonique, *i.e.* une composante normée t associée au tableau concaténé X définie par $t = Xw$, liée de manière optimale aux différents tableaux $(X_1, \ldots, X_K)$. Pour cela, t est projetée sur chaque sous-espace engendré par les variables de X_k. L'ACG recherche à maximiser en moyenne les variances restituées par ces projections, ce qui revient, pour chaque solution d'ordre $h = (1, \ldots, H)$, à maximiser le critère (1).

$$\sum_k var(P_{X_k}t^{(h)}) = \frac{1}{N}t^{(h)'}\sum_k P_{X_k}t^{(h)} \tag{1}$$

La solution de ce problème revient à considérer $(t^{(1)}, \ldots, t^{(H)})$ comme les vecteurs propres normés successifs de la matrice $\sum_k P_{X_k}$. Par la suite, K composantes partielles associées aux tableaux X_k peuvent être exhibées : $t_k^{(h)} = P_{X_k}t^{(h)}/||P_{X_k}t^{(h)}||$.

Soit à présent un tableau de variables quantitatives X partitionné en K blocs $X = [X_1|\ldots|X_K]$ et un tableau Y contenant Q variables quantitatives à expliquer, mesurées sur les mêmes N individus. Nous cherchons désormais une composante u, contrainte à être la combinaison linéaire des variables Y, définie par $u = Yv$. Cette composante est

projetée sur chaque sous-espace engendré par les variables de X_k. Nous souhaitons qu'en moyenne, les variances restituées par ces projections soient les plus grandes possibles. Ainsi, la solution d'ordre un revient à maximiser le critère (2).

$$\sum_k var(P_{X_k}u^{(1)}) = \frac{1}{N}u^{(1)'}\sum_k P_{X_k}u^{(1)} = \frac{1}{N}\sum_k v^{(1)'}Y'\sum_k P_{X_k}Yv^{(1)} \quad (2)$$

En adoptant la contrainte de norme $||v^{(1)}|| = 1$, il s'ensuit que $v^{(1)}$ est le premier vecteur propre de la matrice $Y'\sum_k P_{X_k}Y$, ce qui donne la solution de l'$ACPVI$ multibloc (Bougeard *et al.*, 2007). Des composantes partielles associées à chaque tableau X_k sont exhibées par $t_k^{(1)} = P_{X_k}u^{(1)}/||P_{X_k}u^{(1)}||$. Une composante globale associée au tableau concaténé X est donnée par $t^{(1)} = \sum_k a_k^{(1)}t_k^{(1)}$ avec $a_k^{(1)} = ||P_{X_k}u^{(1)}||/\sqrt{\sum_l ||P_{X_l}u^{(1)}||^2} = cov(u,t_k)/\sqrt{\sum_l cov^2(u,t_l)}$ (Bougeard *et al.*, 2007). Ainsi, les coefficients a_k reflètent le lien entre le tableau Y et les tableaux X_k. Il faut souligner que la composante globale t joue un rôle important, aussi bien pour la description des données que pour la prédiction. Il est préconisé de déterminer la solution d'ordre suivant en effectuant la même démarche et en remplaçant les tableaux X_k et Y par leurs résidus respectifs de la régression sur la première composante globale $t^{(1)}$. Cette procédure est répétée plusieurs fois pour obtenir des composantes globales $(t^{(1)},\ldots,t^{(H)})$, ainsi que les composantes partielles associées. Les composantes globales ainsi obtenues peuvent servir à des fins de prédiction, en régressant les variables Y sur celles-ci. Cette procédure de déflation conduit à l'obtention de composantes globales, résumés de l'ensemble des variables explicatives, mutuellement orthogonales, ce qui améliore la qualité de prédiction du modèle (Westerhuis et Smilde, 2001). Ceci constitue une différence majeure entre l'$ACPVI$ multibloc et l'$ACGTR$ (Kissita, 2003), qui procède par déflation de chaque tableau X_k par rapport à la composante partielle t_k qui lui est associée.

3 ACPVI multibloc qualitative

Soit la variable à expliquer y, une variable qualitative à Q modalités, mesurée sur N individus, codée en un tableau Y de taille $(N \times Q)$ contenant les indicatrices de ses modalités. Nous supposons que le tableau Y est standardisé : $\tilde{Y} = Y(Y'Y)^{-1/2} = YD_Y^{-1/2}$. Ce choix est désormais courant dans le cadre du traitement des variables qualitatives et permet de limiter l'impact de la taille des groupes qui peut être différente selon la modalité. Soit le tableau des variables qualitatives explicatives, constitué de K variables $(x_1,\ldots,x_K)$ mesurées sur les mêmes N individus. Chacune de ces variables x_k, comportant P_k modalités, est codée en un tableau X_k de taille $(N \times P_k)$. Le tableau concaténé est défini par $X = [X_1|\ldots|X_K]$.

Pour résoudre cette problématique, nous cherchons une composante $u^{(1)}$, associée au tableau $\tilde{Y}$ définie par $u^{(1)} = YD_Y^{-1/2}v^{(1)}$. Cette composante est projetée sur chaque

sous-espace engendré par les tableaux X_k. Nous souhaitons qu'en moyenne, les variances restituées par ces projections soient les plus grandes possibles. Ceci revient, pour la solution d'ordre un, à maximiser le critère (3).

$$\sum_k var(P_{X_k}u^{(1)}) = \frac{1}{N}u^{(1)'}\sum_k P_{X_k}u^{(1)} = \frac{1}{N}v^{(1)'}D_Y^{-1/2}Y'\sum_k P_{X_k}D_Y^{-1/2}Yv^{(1)} \qquad (3)$$

En adoptant la contrainte de norme $||v^{(1)}|| = 1$, il s'ensuit que $v^{(1)}$ est le premier vecteur propre de la matrice $D_Y^{-1/2}Y'\sum_k P_{X_k}D_Y^{-1/2}Y$ associé à la plus grande valeur propre $\lambda^{(1)}$. Cette méthode constitue une extension qualitative de l'$ACPVI$ multibloc. De l'équation (3), il découle :

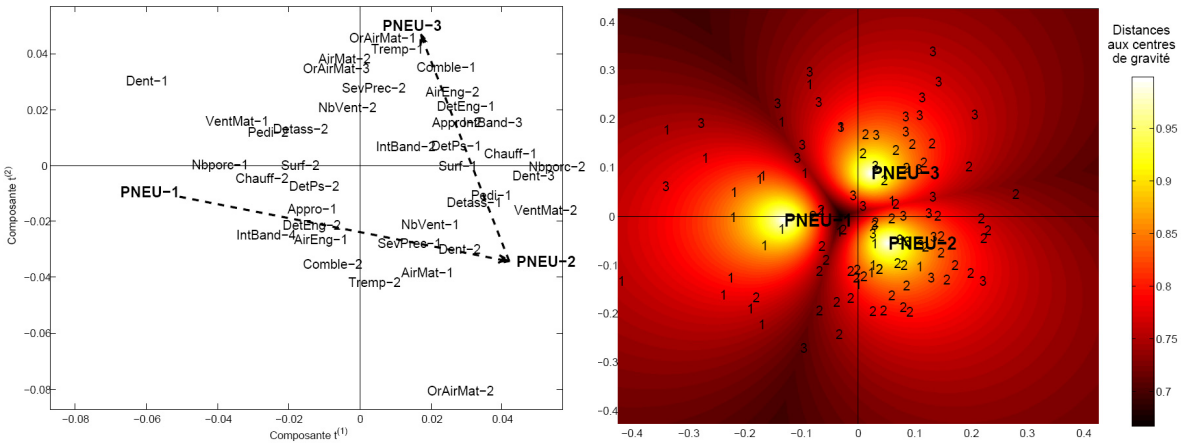
$$\begin{aligned} D_Y^{-1/2}Y'\sum_k P_{X_k}D_Y^{-1/2}Yv^{(1)} &= \lambda^{(1)}v^{(1)} \\ \Longleftrightarrow YD_Y^{-1}Y'\sum_k P_{X_k}D_Y^{-1/2}Yv^{(1)} &= \lambda^{(1)}YD_Y^{-1/2}v^{(1)} \\ \Longleftrightarrow P_Y\sum_k P_{X_k}u^{(1)} &= \lambda^{(1)}u^{(1)} \end{aligned}$$

Il s'ensuit que $u^{(1)}$ est le premier vecteur propre de la matrice $P_Y\sum_k P_{X_k}$. Cette solution revient à la solution d'ordre un de l'Analyse des Correspondances Multiples avec un Tableau de Référence (Kissita, 2003). Des composantes partielles associées à chaque tableau X_k peuvent être exhibées par $t_k^{(1)} = P_{X_k}u^{(1)}/||P_{X_k}u^{(1)}||$. Une composante globale associée au tableau concaténé X peut aussi être donnée par $t^{(1)} = \sum_k a_k^{(1)}t_k^{(1)}$ avec $a_k^{(1)} = ||P_{X_k}u^{(1)}||/\sqrt{\sum_l ||P_{X_l}u^{(1)}||^2}$. Nous choisissons, comme dans le cadre quantitatif présenté dans le paragraphe 2, de déterminer les solutions d'ordre suivant en remplaçant les tableaux X_k et Y par leurs résidus respectifs de la régression sur la première composante globale $t^{(1)}$, et en répétant plusieurs fois cette procédure pour obtenir les composantes suivantes $(t^{(1)}, \ldots, t^{(H)})$.

La qualité de prédiction de la méthode, ainsi que le choix du nombre de composantes à retenir dans le modèle, sont évalués par la qualité de la règle de classement issue du modèle. Les techniques de ré-échantillonnage offrent une solution qui permet d'évaluer le taux apparent d'individus bien classés en se basant sur l'échantillon de calibration, mais aussi le taux théorique d'individus bien classés calculé sur l'échantillon de validation. La règle d'affectation utilisée est basée sur la projection des individus sur l'espace défini par les composantes t, et sur le calcul, dans cet espace comportant $h = (1, \ldots, H)$ composantes, de leurs distances aux centres de gravité des classes. Chaque individu est affecté à la classe correspondant à la plus petite distance.

4 Application

Les données d'épidémiologie animale sont issues d'une enquête analytique portant sur les pathologies respiratoires du porc à l'engrais. Le jeu de données qui en est issu, orienté vers l'étude des facteurs associés à la pneumonie, comporte 105 élevages sur lesquels sont mesurées, après sélection, 19 variables qualitatives, soit 18 variables explicatives et une variable à expliquer. La variable à expliquer ($PNEU$), qui comporte 3 modalités, mesure la sévérité de la pneumonie du lot. Les modalités des variables $(x_1, \ldots, x_K)$ et y peuvent être représentées sur le plan des composantes $t^{(h)}$, qui sont par construction mutuellement orthogonales. La Figure 1 illustre cette représentation pour les deux premières dimensions.



(a) Plan des variables

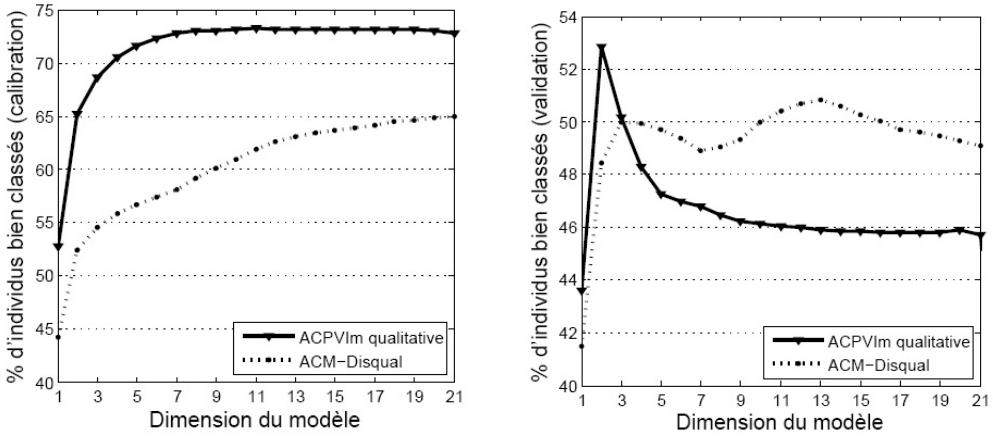
(b) Plan des individus

Figure 1: Représentation factorielle sur le plan des composantes $t^{(1)}$ et $t^{(2)}$.

Le nombre optimal de composantes à retenir dans le modèle de prédiction de y par les variables $(x_1, \ldots, x_K)$ est évalué par validation croisée. Une comparaison à la méthode *Disqual* (Saporta, 1975) est proposée. La Figure 2 illustre l'évolution du taux apparent d'individus bien classés (calibration) et du taux théorique (validation) selon le nombre de composantes introduites dans le modèle.

5 Conclusion et perspectives

Les méthodes étudiées se placent dans le cadre de l'explication d'une variable qualitative y par un ensemble de variables qualitatives $(x_1, \ldots, x_K)$. Nous appliquons l'$ACPVI$ multi-bloc au cadre qualitatif où chaque bloc est constitué par l'ensemble des modalités d'une variable qualitative codée de façon disjonctive. Ainsi, l'application de méthodes multi-blocs, initialement développées pour le traitement de données quantitatives organisées en $(K + 1)$ tableaux, ouvre sur de nouvelles voies de recherche pour les problématiques de



(a) Qualité de modélisation (b) Qualité de prédiction

Figure 2: Evolution de la qualité de modélisation et de prédiction du modèle en fonction du nombre de composantes $(t^{(1)}, \ldots, t^{(H)})$ introduites.

discrimination qualitative. L'avantage direct de ce type d'application est que l'ensemble des modalités d'une variable constitue un bloc, ce qui permet une prise en compte de manière appropriée de la structure des données et procure des aides à l'interprétation de nature à aider l'utilisateur, *i.e.* composantes globales, composantes par blocs, coefficients reflétant l'importance de chaque bloc dans la discrimination. Nous montrons que la solution d'ordre un de cette méthode coïncide avec celle de l'Analyse des Correspondances Multiples avec un Tableau de Référence (Kissita, 2003). Les solutions d'ordre suivant permettent d'optimiser l'aspect prédictif de la méthode. Par ailleurs, l'$ACPVI$ multibloc qualitative peut être étendue au cas de plusieurs tableaux Y à expliquer. Cette extension permet d'expliquer simultanément plusieurs variables qualitatives. Dans le cadre des données d'épidémiologie par exemple, cela permettrait d'expliquer une maladie caractérisée par plusieurs variables ou son évolution évoluant au cours du temps.

Bibliographie

[1] Bougeard, S., Hanafi M. et Qannari, E. M. (2007) ACPVI multibloc. Application à des données d'épidémiologie animale. JSFdS, 148: 77–94.
[2] Carroll, J. D. (1968) A generalization of canonical correlation analysis to three or more sets of variables. *76th annual convention of the Am. psychological association*, 227–228.
[3] Kissita, G. (2003) Les analyses canoniques généralisées avec tableau de référence généralisé : éléments théoriques et appliqués. Thèse de l'Université Paris Dauphine IX.
[4] Saporta, G. (1975) Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de l'Université Paris VI.
[5] Westerhuis, J. A. et Smilde, A. K. (2001) Deflation in multiblock PLS. Journal of Chemometrics, 15: 485–493.